

10/039,849

日 本 国 特 許 庁  
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日      2 0 0 1 年   9 月   4 日  
Date of Application:

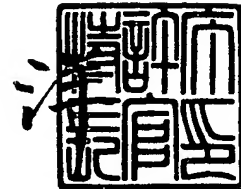
出 願 番 号      特 願 2 0 0 1 - 2 6 6 6 2 9  
Application Number:  
[ J P 2 0 0 1 - 2 6 6 6 2 9 ]  
ST. 10/C):

願      人      株式会社日立製作所  
Applicant(s):

2 0 0 4 年 1 2 月 2 2 日

特許庁長官  
Commissioner,  
Japan Patent Office

小 川



CERTIFIED COPY OF  
PRIORITY DOCUMENT

出証番号   出証特 2 0 0 4 - 3 1 1 7 3 1 5

BEST AVAILABLE COPY

【書類名】 特許願  
【整理番号】 K01000441A  
【あて先】 特許庁長官殿  
【国際特許分類】 G06F 3/06

## 【発明者】

【住所又は居所】 神奈川県小田原市中里 3 2 2 番地 2 号 株式会社日立製作所 S A N ソリューション事業部内

【氏名】 日野 直樹

## 【発明者】

【住所又は居所】 神奈川県小田原市中里 3 2 2 番地 2 号 株式会社日立製作所 R A I D システム事業部内

【氏名】 占部 喜一郎

## 【発明者】

【住所又は居所】 神奈川県小田原市中里 3 2 2 番地 2 号 株式会社日立製作所 R A I D システム事業部内

【氏名】 中野 俊夫

## 【特許出願人】

【識別番号】 000005108

【氏名又は名称】 株式会社日立製作所

## 【代理人】

【識別番号】 100075096

## 【弁理士】

【氏名又は名称】 作田 康夫

## 【手数料の表示】

【予納台帳番号】 013088

【納付金額】 21,000円

## 【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1  
【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 データ格納システム

【特許請求の範囲】

【特許請求の範囲】

【請求項 1】

分散して設置されたデータを格納する第一の記憶装置と第二の記憶装置間でリモートコピー機能により作成されたペアボリュームを、2 台以上の上位装置から共用するデータ格納システムにおいて、1 台の前記上位装置が前記ペアボリュームを排他的に占有し、他の前記上位装置からの更新要求を拒否する制御用ペアボリュームと、データ格納用ペアボリュームから成り、前記制御用ペアボリュームに 2 台以上の前記上位装置から更新要求があった場合に、前記制御用ペアボリュームに最も早くアクセスを行った前記上位装置が前記制御用ペアボリュームを排他的に使用する権利を獲得するデータ格納システム。

【請求項 2】

請求項 1 記載のデータ格納システムにおいて、ペアボリュームの属性及び、前記ペアボリュームの属性を遷移させるペアボリューム制御コマンドを発行する、ペアボリューム制御用ソフトウェアを使用することで、前記制御用ペアボリュームに対して更新要求を発行できる上位装置を排他的に決定するデータ格納システム。

【請求項 3】

請求項 1 または請求項 2 のデータ格納システムにおいて、上位装置上の汎用アプリケーションと連携して動作し、前記記憶装置に対し前記ペアボリューム制御コマンドを発行する前記ペアボリューム制御用ソフトウェアを使用する事で、前記制御用ペアボリュームに対して更新要求を発行できる上位装置を排他的に決定するデータ格納システム。

【請求項 4】

2 台以上の上位装置と、2 台の記憶装置がデータ転送インターフェース及びネットワークインターフェースで接続されるクラスタリングシステムにおいて、前記上位装置もしくはネットワークインターフェースの障害発生時に前記上位装置

の専用使用権を一意に決定するための前記記憶装置内の判定用ボリュームをリモートコピー機能で二重化する、データ格納システム。

**【請求項 5】**

2 台以上の上位装置と、2 台の記憶装置がデータ転送インターフェース及びネットワークインターフェースで接続されるクラスタリングシステムにおいて、前記上位装置もしくはネットワークインターフェースの障害発生時に前記上位装置の専用使用権を一意に決定するための前記記憶装置内の判定用ボリュームをリモートコピー機能で二重化し、複数の前記上位装置から前記判定用ペアボリュームを 1 つのボリュームとして認識するクラスタリングシステムを実現するデータ格納システム。

**【請求項 6】**

分散して設置されたデータを格納する第一の記憶装置と第二の記憶装置間でリモートコピー機能により作成されたペアボリュームを、2 台以上の上位装置から共用するデータ格納システムにおいて、ペアボリュームの専有使用を目的として前記上位装置から発行された S C S I リザーブコマンドを前記ペアボリューム制御用ソフトウェアを介して、前記記憶装置内のボリュームに発行し、前記 S C S I リザーブコマンドに対する前記ボリュームから正常終了のレスポンスを前記ペアボリューム制御用ソフトウェアがうけて、ペアボリューム制御コマンドを前記ボリュームに発行し、前記ボリュームの属性の確認及び、前記属性に応じて、前記ボリュームの前記属性を遷移させる前記コマンドを発行し、正常終了した場合のみ、前記 S C S I リザーブコマンドに対する正常終了を前記上位装置に応答し、前記ボリュームに対して更新要求を発行できる上位装置を排他的に決定するデータ格納システム。

**【発明の詳細な説明】**

**【0 0 0 1】**

**【発明の属する技術分野】**

本発明は分散して設置された 2 台以上の上位装置と 2 台の記憶装置から構成されるデータ格納技術に関する。

**【0 0 0 2】**

更に具体的には分散して設置された2台のリモートコピー機能を持つ記憶装置間で作成された2重化ボリュームの排他的制御を行うデータ格納技術に関する。

#### 【0 0 0 3】

##### 【従来の技術】

システムを止めることなく24時間サービスを継続させるためのさまざまなIT技術がいま、特にEビジネス化を推進する企業の関心を集めている。その中で高可用性を実現する技術の一つがクラスタリングである。複数台のシステムを組み合わせ、一部システムの障害時にもシステム全体が停止せず稼動できるように構築することで、データベース業務を行うシステム等で主に使われている。例えば、小規模システムの例では数台のサーバーで1つの共有ディスクを利用する構成をとり、サーバーと共有ディスク間はSCSIインターフェースを用い、この構成で数台のサーバーが扱っているデータを共有ディスクに置いておけば、一台のサーバーがダウンしても、他の一台のサーバーが共有ディスクのデータを引き継ぎ、処理を継続することができる。これをフェイルオーバーという。このようにフェイルオーバーを行うクラスタリングシステムでは、クライアントからはただ一つのサーバーを引き続き利用しているように業務を行える。

#### 【0 0 0 4】

また、MicroSoft社のクラスタリング製品であるMSCS（マイクロソフト・クラスタ・サーバ）ではクラスタリングシステム内でフェイルオーバーを行う上位装置を一意に決定する為に、クラスタ構成情報を管理するクォーラムリソースと呼ばれる一つの判定用ボリュームを持つ。この判定用ボリュームに対して排他制御を行うことで、そのデータの一貫性を維持できる。

#### 【0 0 0 5】

すなわち、複数の異なるプロセスが同時に一つのリソースに対して重複処理をしてしまうことにより、データの一貫性が損なわれるという問題を回避している。

#### 【0 0 0 6】

次に、2台以上の上位装置と、2台の記憶装置がSCSIインターフェースによって接続される場合のクラスタリングシステムを例として説明する。クラスタ

リングシステムでは、各々の上位装置間をネットワークインターフェースで接続し、各上位装置は互いにメッセージを発信してハートビート通信を行う事で、お互いの稼動状況を監視する。

#### 【0 0 0 7】

さらに、クラスタリングシステムは上位装置の障害又はネットワークインターフェースの障害が発生し、ハートビート通信が不可能になったことを検出して、個々の上位装置が他の上位装置のリソース、アプリケーション及びサービスの引き継ぎを行うかどうかの判断を行う。その判断を行う唯一の決定要素としてクラスタ構成情報を管理する 1 つの判定用ボリュームを記憶装置内に設ける。

#### 【0 0 0 8】

一般に S C S I インターフェースを使用するシステムにおいては、複数のホストのうちひとつのホストがひとつのターゲット（例えば、磁気ディスクドライブ）を排他制御（専用使用）をする場合、最初に S C S I リザーブコマンドを発行することにより、そのターゲットをリザーブでき専用使用が可能となる。

#### 【0 0 0 9】

上位装置の障害又はネットワークの障害等が発生し、ハートビート通信が不可能になった場合、各々の上位装置はクラスタリングシステム内での専用使用権を獲得する為に、判定用ボリュームに対して S C S I リザーブコマンドを発行し、専用使用権を手に入れるを試みる。判定用ボリュームへの S C S I リザーブコマンドの実行が成功した上位装置は他の上位装置のアプリケーション、サービス及びディスクボリューム等のリソースを全て引き継ぎ、フェイルオーバーを実行する。一方、判定用ボリュームに対する S C S I リザーブコマンド実行が失敗した上位装置はその上位装置上で実行していたアプリケーション及びサービスを全て停止する。

#### 【0 0 1 0】

このように、クラスタリングシステムは、一台の上位装置がクラスタ構成情報を管理する判定用ボリュームを専用使用することで実現されている。

#### 【0 0 1 1】

【発明が解決しようとする課題】

特に大量のデータをリアルタイムで処理する必要がある、大規模なオンライントランザクションシステムにおいてはシステムを止めないために、サーバーをはじめディスク装置や電源などを多重化し、一つが故障しても他方が処理を引き継ぐ、フェイルオーバ機能をサポートしたクラスタリングシステムが必要になる。

#### 【0012】

従来技術によれば、クラスタリングシステムにおける判定用ボリュームは全ての上位装置がアクセスすることができる唯一、1つのディスクボリュームによって構成されている。

#### 【0013】

しかし、複数の上位装置が分散配置されるクラスタリングシステムにおいては、1つの判定用ボリュームを全ての上位装置がアクセスできるように設置するには、上位装置と判定用ボリュームを有する記憶装置間のインターフェース、例えばSCSIインターフェース等、接続用のケーブルの接続距離の制限がありシステム構築が出来ない場合があるという問題があった。

#### 【0014】

これを解決する手段として、災害バックアップシステムや遠隔地へのデータ送信システムを構築できるリモートコピー機能をクラスタリングシステムに用いることが考えられる。

#### 【0015】

リモートコピー機能では上位装置を経由せずに（ホストフリー），記憶装置間のリモートコピーを実現する。リモート側への更新データ反映順序を守るため、リモート側のコピーデータの論理的ー貫性の保証が可能である。ここで、リモートコピー機能により分散配置された記憶装置間で二重化されたペアボリューム（それぞれを正ボリューム及び副ボリュームと呼び、正ボリュームを有する記憶装置及び副ボリュームを有する記憶装置がある）を作成し、ペアボリュームの正ボリュームもしくは副ボリュームは同一のデータを持つことが保証されているので、ある一台の上位装置のみが正ボリュームと副ボリュームのどちらかを判定用ボリュームとして専用使用する上位装置を決定する方法が考えられる。

#### 【0016】



しかし、分散配置された各システムはそれぞれ、そのシステム内でそれぞれの SCS I インターフェースを使用しているシステムゆえ、各上位装置はそれぞれのシステム内で SCS I リザーブコマンドを発行し、実行が成功した上位装置がボリュームを排他制御できることになり、正ボリュームを専用使用する上位装置及び副ボリュームを専用使用する2台の上位装置が存在することになる。

#### 【0017】

すなわち、正ボリュームもしくは副ボリュームを排他的に使用する目的で SCS I リザーブコマンドを発行しただけでは、正ボリュームもしくは副ボリュームは別々にリザーブすることができる為に、判定用ボリュームとして排他使用ができないという問題があった。

#### 【0018】

##### 【課題を解決するための手段】

前述に示す通り、リモートコピー機能を用いて判定用ボリュームを設定する場合、ペアを構成する正ボリューム及び副ボリュームは各上位装置から別々のボリュームとして認識され、各上位装置が正ボリューム及び副ボリュームに対し、SCS I リザーブコマンドを発行した場合、正ボリューム及び副ボリュームを個々にリザーブすることができ、結果として複数の上位装置が同じアプリケーション及びサービスを同時に稼動してしまい、クラスタリングシステムとしての機能をはたさないことになる。

#### 【0019】

従って、リモートコピー機能をもつ2台の記憶装置間で作成されたペアボリュームを、複数の上位装置からなるクラスタリングシステムにおいて1つの判定用ボリュームとして認識させ、クラスタリングシステム内の複数の上位装置間で当該ペアボリュームに対し更新要求を発行できる上位装置を排他的に決定する手法を実現する為に、本発明では上位装置上のクラスタソフトウェアから発行される SCS I コマンドのリザーブコマンドを、ペアボリュームの「状態」(属性)を操作するペアボリューム制御用ソフトウェアが一旦受け、ペアボリューム制御用ソフトウェアが発行するペアボリューム制御コマンドと組み合わせて記憶装置に発行し、そのコマンドに対する記憶装置のレスポンスが、最初にリザーブコマン

ドを発行した上位装置のみ正常終了とすることで、ペアボリュームを排他的使用する上位装置を一意に決定することが可能となるデータ格納システムを提案する。

### 【0020】

#### 【発明の実施の形態】

以下、本発明の一実施例を図面を用いて説明する。尚、クラスタリングシステムを利用した実施例としたが、本発明の適用範囲はクラスタリングシステムに限ったものではない。

また、2台の上位装置からなるクラスタリングシステムを実施例として用いたが、上位装置は3台以上あっても構わない。

### 【0021】

図1は本発明の一実施形態である。図1はそれぞれ上位装置と記憶装置からなる2つのサイト100、101をネットワークインターフェースによって接続し、クラスタリングシステムを実現する構成図である。10、11はCPUを持つ上位装置であり、上位装置10、11上ではクラスタソフトウェア12、13及び本発明で提案するペアボリューム制御用ソフトウェア84、85等が動作する。また、上位装置10、11はハートビート用ネットワーク50によって接続され、定期的にお互いの稼動状況を監視する。

### 【0022】

20、21は上位装置10、11からのデータを格納しておく記憶装置であり、データ転送インターフェース30、31を介して接続される。また、記憶装置20、21は複数のボリュームから構成され、上位装置10、11上にインストールされたアプリケーションからのデータ更新及び参照要求を受けることができる。

### 【0023】

さらに、記憶装置20、21はリモートコピー機能を持ち、一方の記憶装置上の正ボリューム（P-VOL）22のミラーをもう一方の記憶装置上の副ボリューム（S-VOL）23に構築することができる。上位装置10で動作するアプリケーションは記憶装置20内の正ボリューム（P-VOL）22に対してデー

タの更新要求を行い、その更新内容がリモートコピー用インターフェース 4 0 を通じて自動的にもう一方の記憶装置内の副ボリューム（S-VOL）2 3 に反映される。

#### 【0 0 2 4】

図 2 はクラスタリングシステムの内部構造を示す。上位装置 1 0, 1 1 はそれぞれクラスタソフトウェア 8 0、8 1、アプリケーション 8 2, 8 3 及びペアボリューム制御用ソフトウェア 8 4, 8 5 から構成される。ペアボリューム制御用ソフトウェア 8 4、8 5 はクラスタソフトウェア 8 0, 8 1 と標準デバイスドライバ 9 0, 9 1 の中間に位置する。

#### 【0 0 2 5】

クラスタソフトウェア 8 0, 8 1 はペアボリューム制御用ソフトウェア 8 4、8 5 が独自に定義したデバイスファイルに対して S C S I コマンドを発行し、それを受け取ったペアボリューム制御用ソフトウェア 8 4、8 5 は、必要に応じてペアボリューム制御コマンドと組み合わせたコマンドを標準デバイスドライバ 9 0, 9 1 経由でクラスタ構成情報を管理する判定用ペアボリューム 6 0, 6 1 へ発行する。

#### 【0 0 2 6】

また、記憶装置 2 0, 2 1 はクラスタの専有使用権を確保する為に構築された判定用ペアボリューム 6 0, 6 1、及びアプリケーション用のデータを格納するデータボリューム 7 0 - 7 3 で構成される。リモートコピー機能によって各ボリュームは二重化されており、リモートコピー機能の仕様により、各アプリケーションは正ボリューム（P-VOL）のみに更新要求を発行する事ができる。

#### 【0 0 2 7】

クラスタソフトウェア 8 0 と 8 1 はネットワークを介してハートビート通信を行うことで定期的に相手側上位装置の稼動状況を監視する。上位装置 1 0 又は 1 1 の障害、又はハートビートネットワーク 5 0（図 1）の障害によりハートビート通信が不可能となった場合、クラスタソフトウェア 8 0、8 1 はペアボリューム制御用ソフトウェア 8 4, 8 5 を介して、記憶装置 2 0, 2 1 内に構築された判定用ペアボリューム 6 0, 6 1 にそれぞれ S C S I リザーブコマンドを発行す

る。

#### 【0 0 2 8】

そして、判定用ペアボリュームをリザーブすることに成功した上位装置はクラスタソフトウェアによりクラスタリングシステム内に登録された全リソース及びアプリケーションの引き継ぎを行う。判定用ペアボリュームをリザーブすることに失敗した他の上位装置はクラスタソフトウェアによりリソースを解放し、アプリケーションを停止する。

#### 【0 0 2 9】

ここで、図 3 に示すペアボリュームの状態遷移について詳細説明する。リモートコピー機能によって生成された各ペアボリュームは一般的に「非ペア状態」、「コピー状態」、「ペア状態」、「サスペンド状態（正常）」、「サスペンド状態（異常）」の状態を持ち、上位装置上にインストールされたペアボリューム制御ソフトウェアがペアボリューム制御コマンドを記憶装置へ発行することで、リモートコピー用インターフェース 4 0 を通じて自動的に各々のペアボリュームの状態を遷移させることができる。

#### 【0 0 3 0】

「非ペア状態」は 2 つのボリューム間にペアが形成されてない状態、「コピー状態」は 2 つのボリューム間で正ボリュームから副ボリュームへコピー実行中でペアボリュームのデータはまだ全て同等ではない状態。コピー完了時、ボリュームは「ペア状態」に変わる

「ペア状態」では、上位装置から正ボリュームへの全ての更新は副ボリュームに対して二重化される。「サスペンド状態（正常）」は正ボリュームに対する更新を副ボリュームに反映しない状態。例えばペアボリューム間で正ボリュームのみ更新をかけたい時、正、副ボリュームを本状態に設定する。

#### 【0 0 3 1】

「サスペンド状態（異常）」は前述と違い、何らかの障害検知で、ペアボリューム間で内容の更新を保持できない時、正、副ボリュームを本状態に設定する。各ペアボリュームの状態を遷移させるペアボリューム制御コマンドは正ボリュームまたは副ボリュームのどちらに接続された上位装置からでも実行可能である。

**【 0 0 3 2 】**

例えば、ペアボリュームが「サスペンド状態（正常）」であった場合、正ボリューム側の上位装置からペア再同期コマンドを発行し「ペア状態」に遷移させる事も可能であるし、副ボリューム側の上位装置からペア再同期コマンドを発行し、「ペア状態」に遷移させることもできる。さらに、副ボリューム側の上位装置からペアスワップ再同期コマンドを発行し、副ボリュームを正ボリュームにスワップした上で「ペア状態」に遷移させることもできる。

**【 0 0 3 3 】**

また、ペアボリューム制御コマンドの実行結果は実行先のペアボリュームの状態に依存するという性質を持つ。以下詳細説明する。

**【 0 0 3 4 】**

図 4 にペアボリューム制御コマンドであるペア再同期コマンド及びペアスワップ再同期コマンド発行時の記憶装置からのレスポンスの例を示す。例えば、実行コマンドとしてペア再同期コマンドを発行した場合、実行先のボリュームが正ボリュームでかつ「サスペンド状態（正常）」でなければ記憶装置からのレスポンスは正常終了とはならない。

**【 0 0 3 5 】**

もし、実行先のボリュームが「コピー状態」もしくは「ペア状態」であった場合はそのペア再同期コマンドは異常終了する。また、ペアスワップ再同期コマンドは実行先のペアボリュームが副ボリュームであり、かつ「サスペンド状態（正常）」でなければ記憶装置からのレスポンスは正常終了とはならない。

**【 0 0 3 6 】**

もし、実行先のボリュームが「コピー状態」もしくは「ペア状態」であった場合はそのコマンドは異常終了する。このペアボリュームの性質を用いて、ペアボリューム制御用ソフトウェアはクラスタソフトウェアから発行された S C S I コマンドを一旦受け取り、本 S C S I コマンドを分析した結果、リザーブコマンド等のペアボリューム間で排他的な処理を必要とする S C S I コマンドであった場合はペアボリューム制御コマンド例えば、ペア再同期コマンドと組み合わせて記憶装置側へ発行する。

**【0037】**

そして、ペア再同期コマンドが記憶装置内で実行され、正常にペアボリューム状態を遷移できた場合は、正常終了ステータスをペアボリューム制御用ソフトウェアに送信する。さらに、正常終了ステータスを受け取ったペアボリューム制御ソフトウェアはクラスタソフトウェアに対して正常終了ステータスを返す。

**【0038】**

正常終了ステータスを受け取ったクラスタソフトウェアはクラスタリングシステム内で引き続き生存することが許可され、フェイルオーバを実行する。一方、1つの上位装置から最初に発行されたペア再同期コマンドが記憶装置上で実行され、ペアボリュームの状態が「ペア状態」に遷移したことによって、他の上位装置から同様のペアボリューム制御コマンドが発行されたとしても図4に示すよう、「ペア状態」に発行されたペアボリューム制御コマンドは全て異常終了となる。

**【0039】**

ペアボリューム制御コマンドの異常終了報告を受けたペアボリューム制御用ソフトウェアはSCSIコマンド要求が失敗したことの報告として、クラスタソフトウェアへ異常終了ステータスを返す。異常終了ステータスを受け取ったクラスタソフトウェアはクラスタリングシステム内での生存が許されず、担当しているアプリケーション及びサービスを全て停止させる。

**【0040】**

図5、6で、本発明で提案するペアボリューム制御用ソフトウェアの動作についてSCSIリザーブコマンドを受けた場合とそれ以外のコマンドをうけた場合について更に詳細に説明する。ペアボリューム制御用ソフトウェア84はクラスタソフトウェア80から発行されたSCSIコマンドがリザーブコマンドであった場合(110)、判定用ペアボリュームの正ボリューム(P-VOL)または副ボリューム(S-VOL)に対し、SCSIリザーブコマンドを発行する(111)。

**【0041】**

SCSIリザーブコマンドに対するボリュームからのレスポンスが正常終了で

あった場合（1 1 2）はボリュームが判定用ペアボリュームの正ボリューム（P-VOL）か副ボリューム（S-VOL）かを調査する為にボリューム制御用ソフトウェア 8 4 は記憶装置に対してボリューム属性確認コマンドを発行する（1 1 3）。判定用ボリュームが正ボリューム（P-VOL）であった場合は更にペア再同期コマンドを発行する（1 1 4）。

#### 【0 0 4 2】

一方、判定用ボリュームが副ボリューム（S-VOL）であった場合はペアスワップ再同期コマンドを発行する（1 1 5）。ここで、ペア再同期コマンドとはサスペンド状態のペアボリュームを再同期しペア状態に遷移させるコマンドである。

#### 【0 0 4 3】

一方、ペアスワップ再同期コマンドとはサスペンド状態のペアボリュームを再同期しペア状態に遷移させ、かつ正ボリュームを副ボリュームに、副ボリュームを正ボリュームにスワップするコマンドである。これにより、副ボリュームであった判定用ボリュームが正ボリュームに遷移する。

#### 【0 0 4 4】

図 4 に示すように、ペア再同期コマンドはコマンド発行配下のボリュームすなわち自ボリュームが正ボリューム（P-VOL）かつサスペンド状態（正常）の場合のみ成功する（1 3 0）。

#### 【0 0 4 5】

一方、ペアスワップ再同期コマンドは自ボリュームが副ボリューム（S-VOL）かつサスペンド状態（正常）の場合のみ成功する（1 3 1）。

#### 【0 0 4 6】

また、図 3 に示すように、ペア再同期コマンドまたはペアスワップ再同期コマンドのどちらかが発行され、実行状態になった場合に、ペアボリュームの状態は「サスペンド状態（正常）」（1 2 0）から「コピー状態」（1 2 1）さらに、コピーが完了すると「ペア状態」（1 2 2）に遷移する。よって、図 4 に示すように、ペアボリュームの状態が「ペア状態」になった後で発行されたペア再同期コマンドまたはペアスワップ再同期コマンドは異常終了することになる。

**【0047】**

このようにペアボリューム制御用ソフトウェアはペア再同期またはペアスワップ再同期コマンドが正常終了した場合に、SCSIリザーブコマンドが成功したことを上位装置へGoodステータス（成功）を返して報告する（117）。すなわちクラスタの専用使用権が取れたことを報告する。

**【0048】**

一方、ペアボリューム制御コマンドが異常終了した場合は、リザーブに失敗した、すなわちクラスタの専用使用権が取れなかったことを報告する為に、上位装置に対して、Reservation Conflictステータス（失敗）を返す（116）。

**【0049】**

判定用ボリュームへのリザーブ要求に対して、Goodステータスを得た上位装置はフェイルオーバを実行し、クラスタ内の全てのリソース、アプリケーション及びサービスを引き継ぐ。

**【0050】**

一方、Reservation Conflictステータスを得た上位装置は自身が管理していた全てのリソースを解放し、かつ自身が管理していた全てのアプリケーション及びサービスを停止する。

**【0051】**

また図6に示すように、クラスタソフトウェア80、81から発行されたSCSIコマンドがリザーブコマンド以外であった場合はペアボリューム制御用ソフトウェアはペアボリューム制御コマンドを付加せず、そのままの状態に判定用ボリュームに当該SCSIコマンドを発行する。

**【0052】**

図7から図13にて本発明をクラスタリングシステムに適用した場合の、システムが正常な状態から障害によりクラスタリングが機能する一連の動作について説明する。まず、最初の状態ではクラスタリングシステムは停止し、判定用ペアボリュームは「サスペンド状態（正常）」を維持している（図7）。

**【0053】**

ここで、ユーザがクラスタリングシステムを起動した場合、各上位装置内のク



ラストソフトウェアはクラスタリングシステム内の専用使用権を獲得する為に、ペアボリューム制御用ソフトウェアに対して S C S I リザーブコマンドを発行する。S C S I リザーブコマンドを受領したペアボリューム制御用ソフトウェアは、判定用ペアボリュームの正ボリューム（判定用（P））または副ボリューム（判定用（S））に対し、S C S I リザーブコマンドを発行し、専用使用権争いを行う。本例では各システム内に一台の上位装置のみ存在するので、上位装置 1 0 , 1 1 が発行する S C S I リザーブコマンドに対する各ボリュームからペアボリューム制御用ソフトウェアに対するレスポンスはそれぞれ正常終了となる。

#### 【 0 0 5 4 】

次に、判定用ペアボリュームの正ボリューム（判定用（P））側に接続するペアボリューム制御用ソフトウェアはペア再同期コマンドを発行する。一方、判定用ペアボリュームの副ボリューム（判定用（S））側に接続するペアボリューム制御用ソフトウェアはペアスワップ再同期コマンドを発行する。

#### 【 0 0 5 5 】

本例では上位装置 1 0 のペア再同期コマンドが上位装置 1 1 のペアスワップ再同期コマンドより先に発行され、クラスタの専用使用権を獲得したとする。専用使用権を獲得した上位装置 1 0 内のクラスタソフトウェアは Write コマンドを判定用ペアボリュームに対して発行し、クラスタ管理情報の初期化を行う（図 8）。

#### 【 0 0 5 6 】

次に、各上位装置内のクラスタソフトウェアはユーザにて予め設定されたデータボリュームをマウントし、そのデータボリュームに対するアプリケーションを起動する（図 9）。

#### 【 0 0 5 7 】

さらに、判定用ペアボリュームに対してはシステムの障害に備え、ペアボリューム制御用ソフトウェアを用いて判定用ペアボリュームを分割することで「サスペンド状態（正常）」に遷移させ（図 1 0）、システムは通常運用に入る（図 1 1）。

#### 【 0 0 5 8 】

通常運用中に上位装置の障害又はネットワークの障害等が発生し、ハートビート通信が不可能になった場合、各々の上位装置内のクラスタソフトウェアはクラスタリングシステム内での専用使用権を獲得する為に、SCSIバスデバイスリセットを発行し、データ転送インターフェース上をリセットした後、ペアボリューム制御用ソフトウェアに対してSCSIリザーブコマンドを発行する。SCSIリザーブコマンドを受領したペアボリューム制御用ソフトウェアは、判定用ペアボリュームの正ボリューム（判定用（P））または副ボリューム（判定用（S））に対し、SCSIリザーブコマンドを発行し、専用使用権争いを行う。

#### 【0059】

本例では各サイト内に一台の上位装置のみ存在するので、上位装置10、11が発行するSCSIリザーブコマンドに対し、各ボリュームからペアボリューム制御用ソフトウェアに対するレスポンスはそれぞれ正常終了となる。次に、判定用ペアボリュームの正ボリューム（判定用（P））側に接続するペアボリューム制御用ソフトウェアはペア再同期コマンドを発行する。

#### 【0060】

一方、判定用ペアボリュームの副ボリューム（判定用（S））側に接続するペアボリューム制御用ソフトウェアはペアスワップ再同期コマンドを発行する。本例では上位装置10のペア再同期コマンドが上位装置11より先に発行され、クラスタの専用使用権を獲得したとする（図12）。

#### 【0061】

専用使用権を獲得した上位装置10内のクラスタソフトウェアは他の上位装置のアプリケーション及びディスクボリューム等のリソースを全て引き継ぎ、フェイルオーバーを実行する。一方、判定用ボリュームに対するSCSIリザーブコマンドが失敗した上位装置はその上位装置上で実行していたアプリケーション及びサービスを全て停止する（図13）。

#### 【0062】

このように、本発明により分散して設置された2台のリモートコピー機能を持つ記憶装置間で作成されたペアボリュームを、1台の上位装置が当該ペアボリュームを排他的に占有し、他の上位装置からの更新要求を拒否することで、ペアボ

リュームを当該複数の上位装置に対して1つのボリュームとして認識させることができ、当該ペアボリュームをクラスタリングシステムの判定用ボリュームに適用することができる。

### 【0 0 6 3】

#### 【発明の効果】

本発明は分散して設置された2台のリモートコピー機能を持つ記憶装置間で作成されたペアボリュームを、2台以上の上位装置から共用する構成において、1台の前記上位装置が前記ペアボリュームを排他的に占有し、他の前記上位装置からの更新要求を拒否することで、前記ペアボリュームを前記複数の上位装置に対して1つのボリュームとして認識させるデータ格納システムを実現する。

#### 【図面の簡単な説明】

##### 【図 1】

本発明の一実施形態を示すハードウェア構成図である。

##### 【図 2】

リモートコピー機能のペアボリュームを判定用ボリュームに使用したクラスタリングシステムの構成図である。

##### 【図 3】

ペアボリュームの状態遷移図である。

##### 【図 4】

ペア再同期コマンド又はペアスワップ再同期コマンド発行する場合の、各ボリューム属性に対する記憶装置からのレスポンスを示す

##### 【図 5】

SCSIリザーブコマンド受領時のペアボリューム制御用ソフトウェアの動作である。

##### 【図 6】

SCSIリザーブコマンド以外受領時のペアボリューム制御用ソフトウェアの動作である。

##### 【図 7】

判定用ペアボリュームを持つクラスタリングシステムの動作を示す。

**【図 8】**

判定用ペアボリュームを持つクラスタリングシステムの動作を示す。

**【図 9】**

判定用ペアボリュームを持つクラスタリングシステムの動作を示す。

**【図 1 0】**

判定用ペアボリュームを持つクラスタリングシステムの動作を示す。

**【図 1 1】**

判定用ペアボリュームを持つクラスタリングシステムの動作を示す。

**【図 1 2】**

判定用ペアボリュームを持つクラスタリングシステムの動作を示す。

**【図 1 3】**

判定用ペアボリュームを持つクラスタリングシステムの動作を示す。

**【符号の説明】**

1 0、1 1 上位装置

1 2、1 3 クラスタソフトウェア

2 0、2 1 記憶装置

2 2 正ボリューム (P-VOL)

2 3 副ボリューム (S-VOL)

3 0、3 1 データ転送用インターフェース

4 0 リモートコピー用インターフェース

5 0 ハートビート用ネットワーク

6 0 判定用ペアボリュームの正ボリューム (P-VOL)

6 1 判定用ボリュームの副ボリューム (S-VOL)

7 0、7 3 データボリュームの正ボリューム (P-VOL)

7 1、7 2 データボリュームの副ボリューム (S-VOL)

8 0、8 1 クラスタソフトウェア

8 2、8 3 アプリケーション

8 4、8 5 ペアボリューム制御用ソフトウェア

9 0、9 1 標準デバイスドライバ

100、101 サイト

110 クラスタソフトウェアのSCSIリザーブコマンド発行

111 ペアボリューム制御用ソフトウェアのSCSIリザーブコマンド発行

112 実行結果の判定

113 ボリューム属性の確認

114 ペア再同期コマンド

115 ペアスワップ再同期コマンド

116 リザーブコマンドの失敗 (Reservation Conflictステータス) 117 リ  
ザーブコマンドの成功 (Goodステータス)

120 サスペンド状態 (正常)

121 コピー状態

122 ペア状態

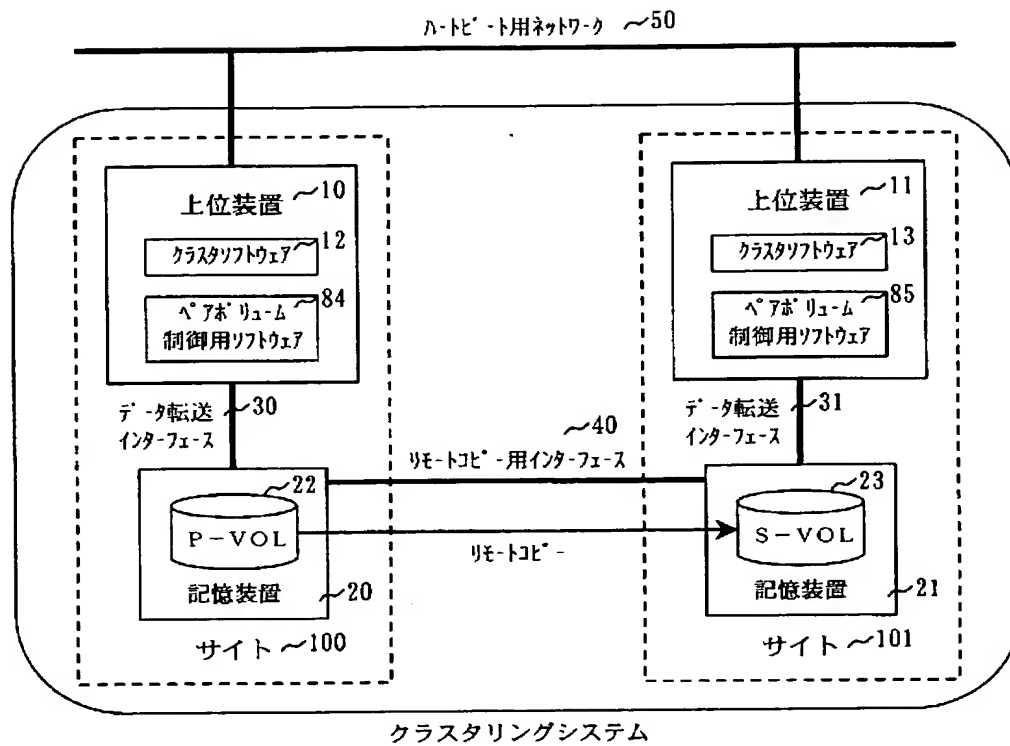
130 ペア再同期コマンドの正常終了

131 ペアスワップ再同期コマンドの正常終了

【書類名】 図面

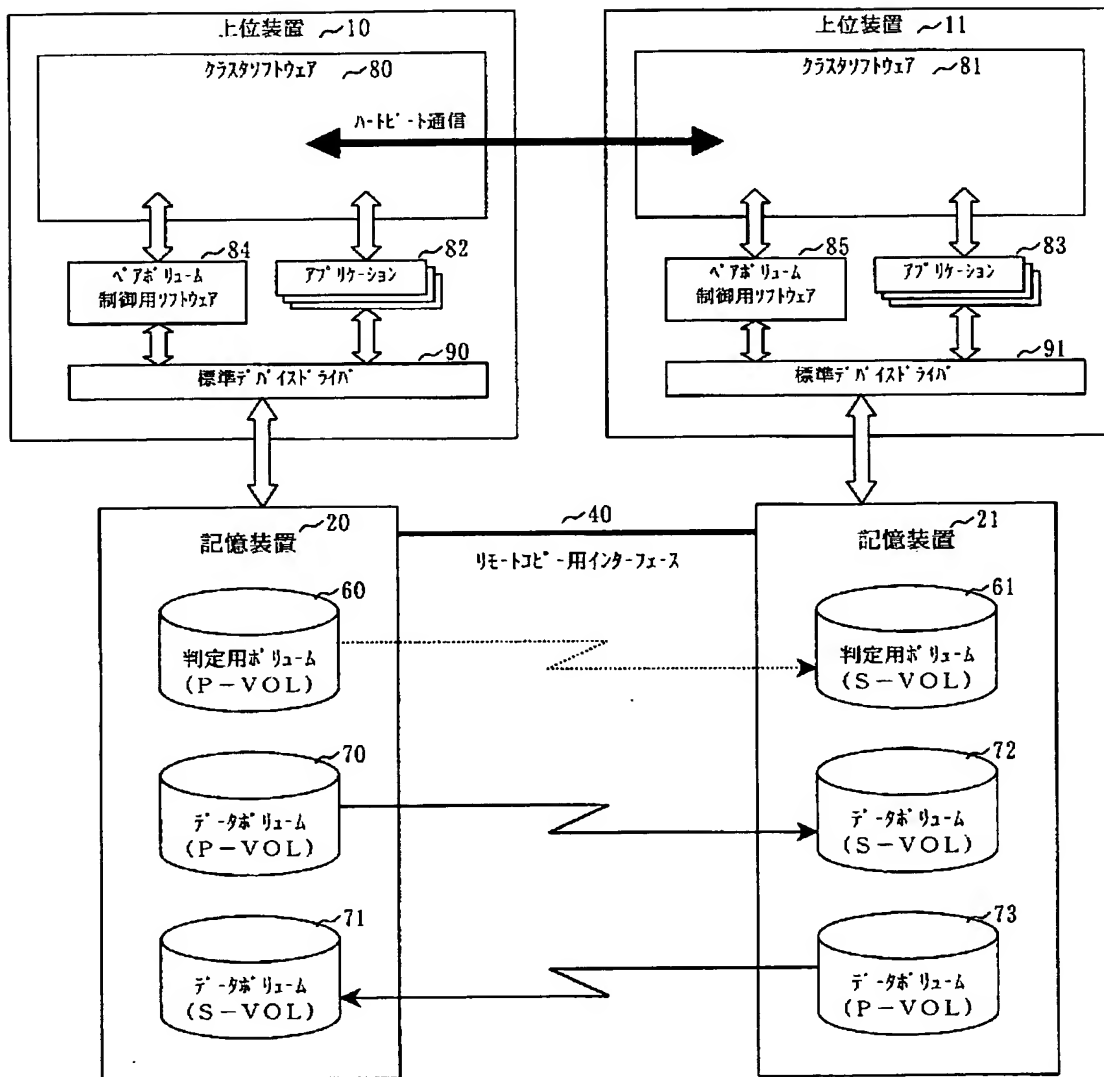
【図 1】

図 1



【図 2】

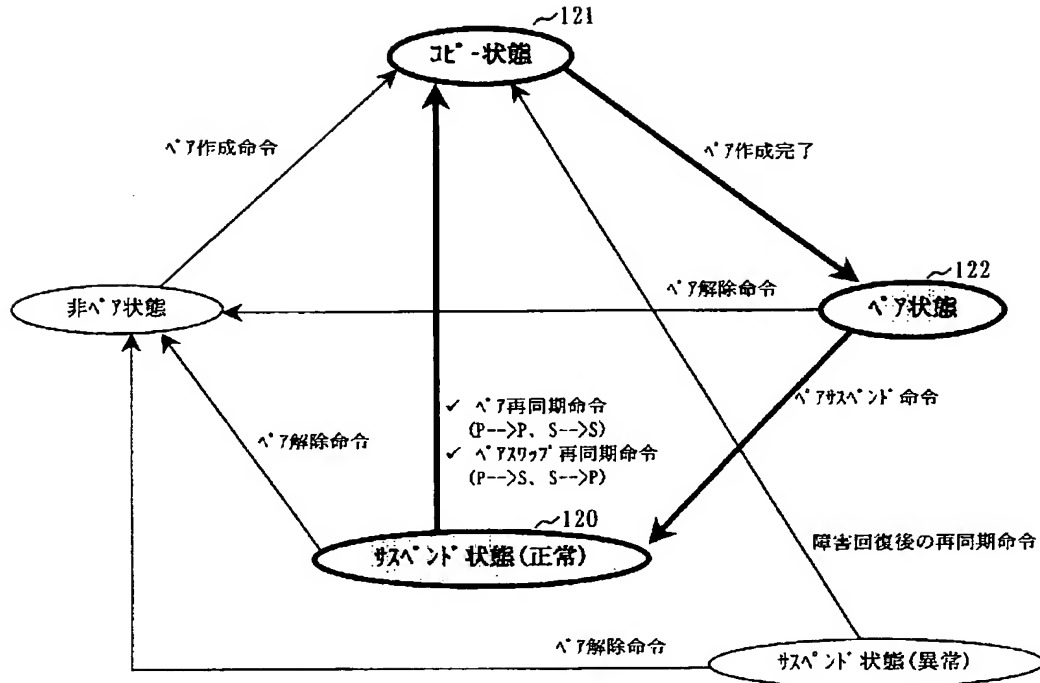
図 2



【図 3】

図 3

ペアボリュームの状態遷移図





【図 4】

図 4

ペア再同期コマンド発行時の記憶装置からのレスポンス

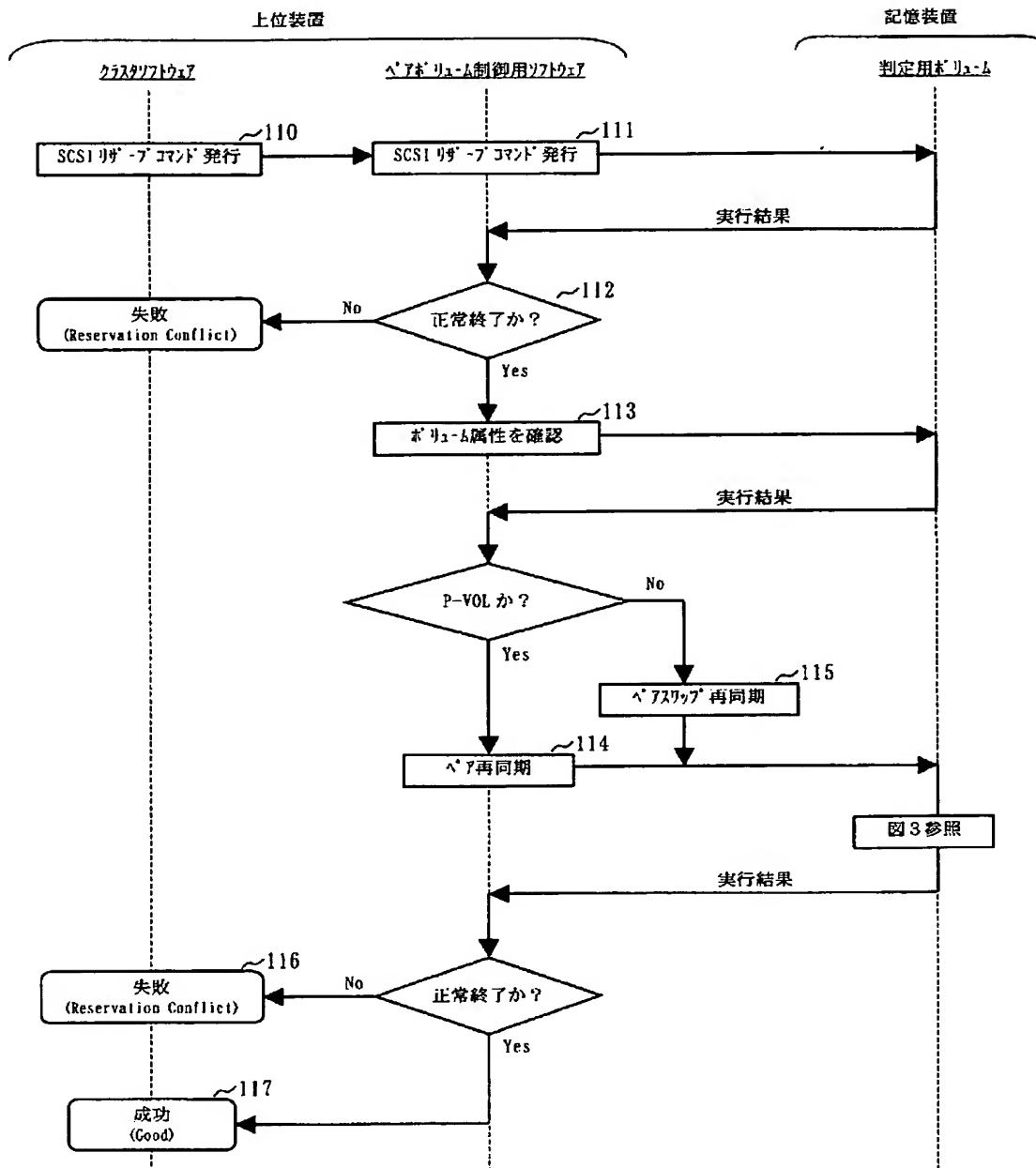
| 実行コマンド    | 自ホストリユームの属性 |             | 記憶装置のレスポンス   |
|-----------|-------------|-------------|--------------|
| ペア再同期     | 非ペアホストリユーム  |             | 異常終了<br>～130 |
|           | P-VOL       | コピ-中        |              |
|           |             | ペア状態        |              |
|           |             | サスペンド状態(正常) |              |
|           |             | サスペンド状態(異常) | 正常終了         |
|           | S-VOL       | コピ-中        | 異常終了         |
|           |             | ペア状態        |              |
|           |             | サスペンド状態(正常) |              |
|           |             | サスペンド状態(異常) |              |
| ペアスワップ再同期 | 非ペアホストリユーム  |             | 異常終了<br>～131 |
|           | P-VOL       | コピ-中        |              |
|           |             | ペア状態        |              |
|           |             | サスペンド状態(正常) |              |
|           |             | サスペンド状態(異常) |              |
|           | S-VOL       | コピ-中        |              |
|           |             | ペア状態        |              |
|           |             | サスペンド状態(正常) | 正常終了         |
|           |             | サスペンド状態(異常) | 異常終了         |

【図 5】

図 5

ペアボリューム制御用ソフトウェアの動作

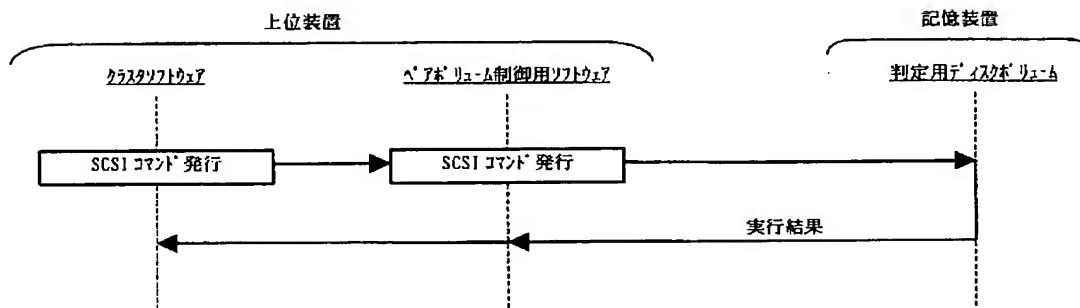
[リザーブコマンド]



【図 6】

図 6

ペアボリューム制御用ソフトウェアの動作  
[リザーブコマンド以外]

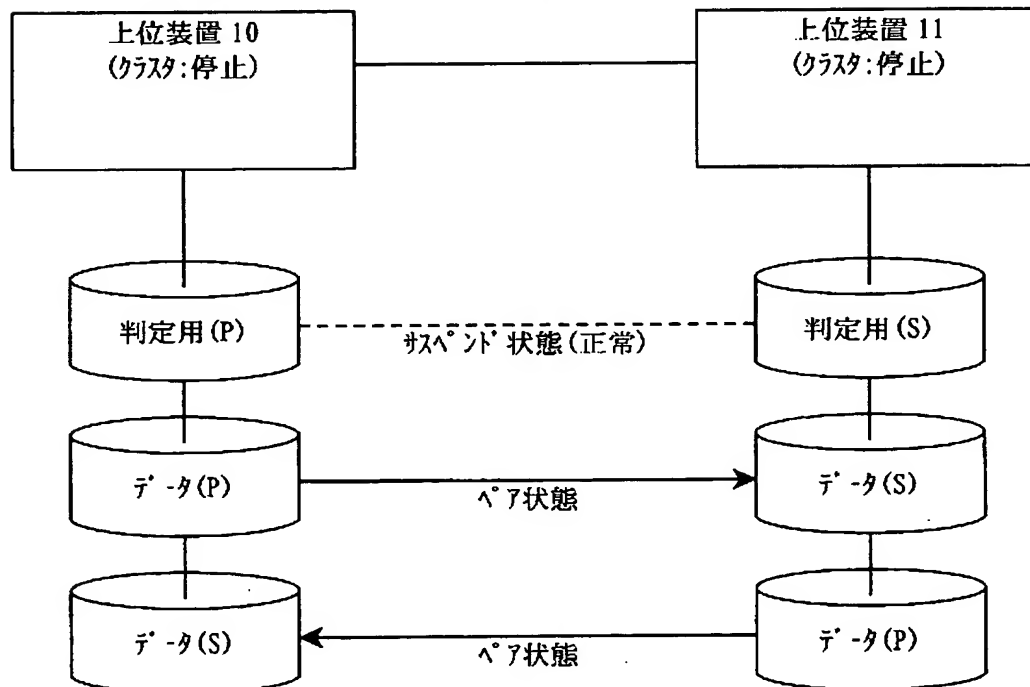


【図 7】

図 7

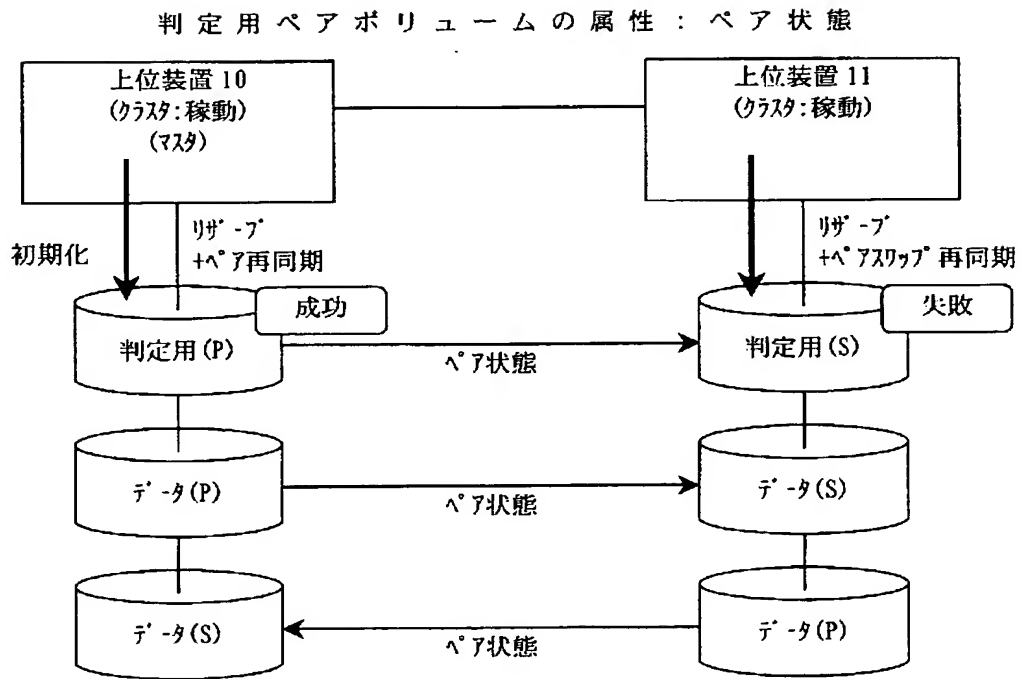
初期状態

判定用ペアボリュームの属性：サスペンド（正常）



【図 8】

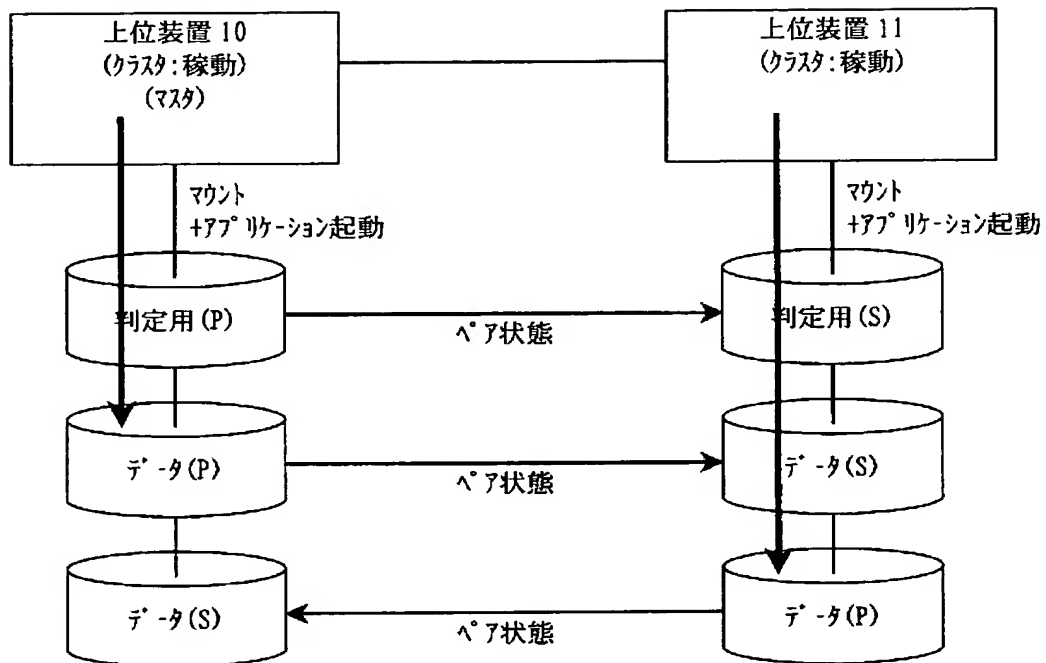
図 8



【図 9】

図 9

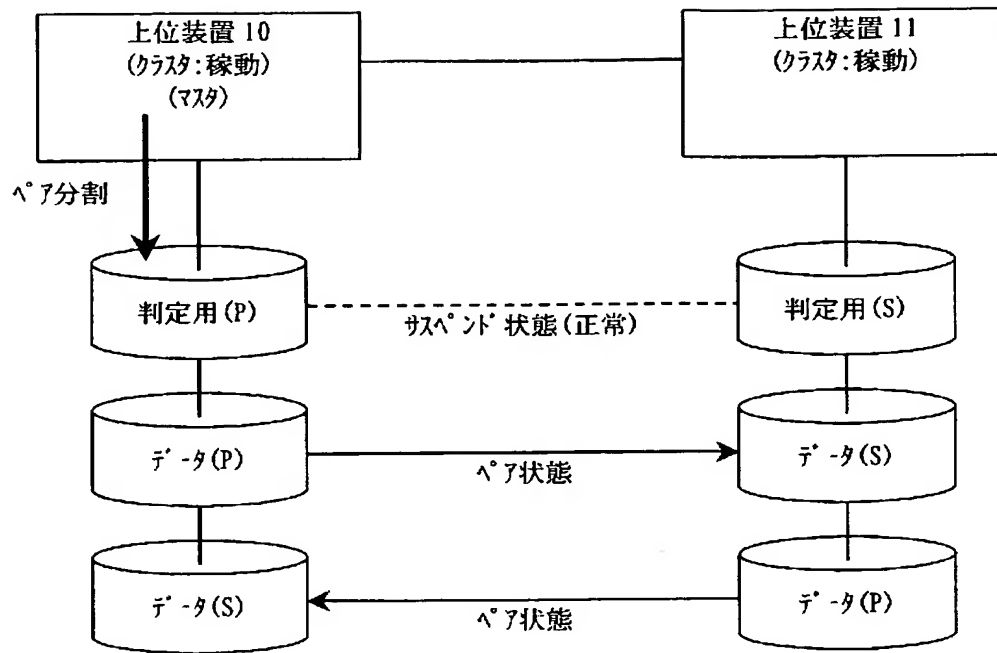
判定用ペアボリュームの属性：ペア状態



【図 10】

図 10

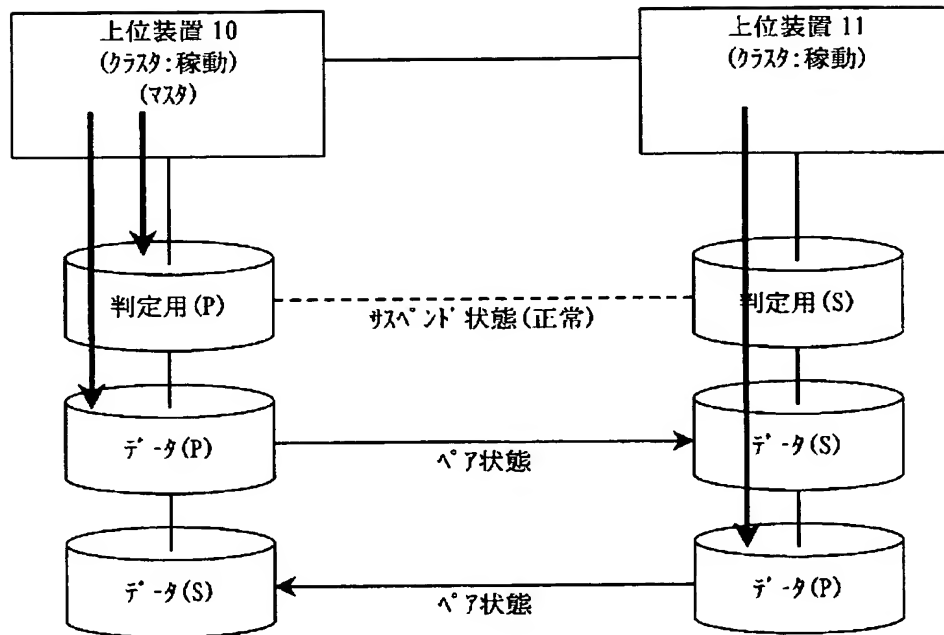
判定用ペアボリュームの属性：サスペンド（正常）



【図 11】

図 11

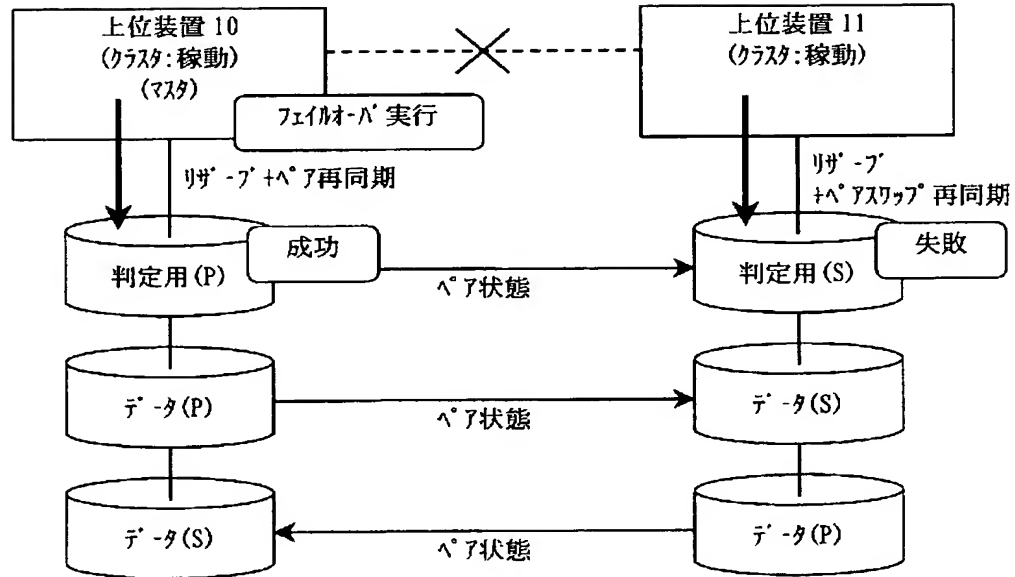
判定用ペアボリュームの属性：サスペンド（正常）



【図 12】

図 12

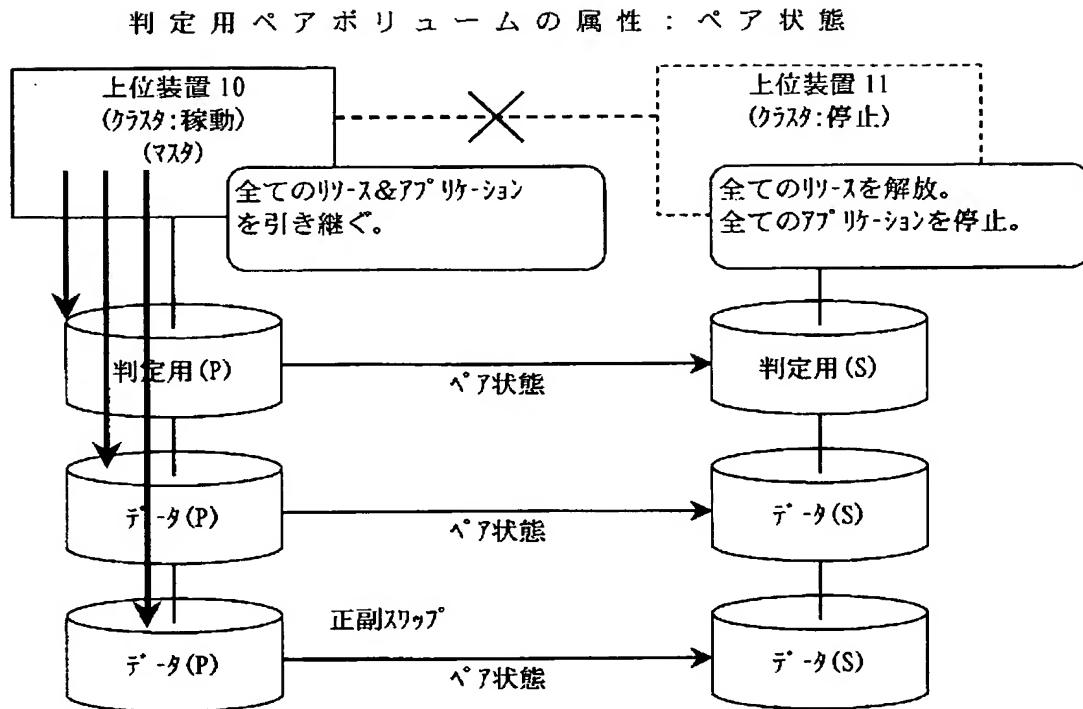
判定用ペアボリュームの属性：ペア状態





【図 13】

図 13



【書類名】 要約書

【要約】

【課題】

リモートコピー機能によって生成されたペアボリュームを、クラスタリングシステムの判定用ボリュームとして使用するデータ格納システムを実現する。

【解決手段】

分散して設置された 2 台のリモートコピー機能を持つ記憶装置 2 0, 2 1 間で作成されたペアボリューム 2 2, 2 3 を、2 台以上の上位装置 1 0, 1 1 から共用する構成において、上位装置 1 0、1 1 上のクラスタソフトウェア 1 2, 1 3 から発行されるリザーブコマンド等の S C S I コマンドを一旦、ペアボリューム制御用ソフトウェア 8 4, 8 5 が受け、当該ペアボリュームが唯一で共通の「状態」となるようにして操作するペアボリューム制御コマンドとを、リザーブコマンド等の S C S I コマンドと組み合わせて発行することにより、ペアボリュームを排他的使用する上位装置を一意に決定する。

【効果】

リモートコピー機能によって生成されたペアボリュームを、クラスタリングシステムの判定用ボリュームとして使用することが可能となる。

【選択図】 図 1

認定・付加情報

|         |                          |
|---------|--------------------------|
| 特許出願の番号 | 特願 2 0 0 1 - 2 6 6 6 2 9 |
| 受付番号    | 5 0 1 0 1 2 9 2 8 4 8    |
| 書類名     | 特許願                      |
| 担当官     | 第七担当上席 0 0 9 6           |
| 作成日     | 平成 1 3 年 9 月 5 日         |

< 認定情報・付加情報 >

|       |             |
|-------|-------------|
| 【提出日】 | 平成13年 9月 4日 |
|-------|-------------|

次頁無

特願 2 0 0 1 - 2 6 6 6 2 9

出 願 人 履 歴 情 報

識別番号 [ 0 0 0 0 0 5 1 0 8 ]

- |          |                        |
|----------|------------------------|
| 1. 変更年月日 | 1 9 9 0 年 8 月 3 1 日    |
| [変更理由]   | 新規登録                   |
| 住 所      | 東京都千代田区神田駿河台 4 丁目 6 番地 |
| 氏 名      | 株式会社日立製作所              |
|          |                        |
| 2. 変更年月日 | 2 0 0 4 年 9 月 8 日      |
| [変更理由]   | 住所変更                   |
| 住 所      | 東京都千代田区丸の内一丁目 6 番 6 号  |
| 氏 名      | 株式会社日立製作所              |